IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 24, 2018, accepted September 19, 2018, date of publication October 1, 2018, date of current version October 31, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2872805

Research on Intelligent Analysis and Depth Fusion of Multi-Source Traffic Data

GUOJIANG SHEN[1], XIAO HAN[1], JUNJIE ZHOU[2], ZHONGYUAN RUAN[1], AND QIHONG PAN[3]

[1]Zhejiang University of Technology, Hangzhou 310023, China
[2]Zhejiang Supcon information Co., Ltd., Hangzhou 310053, China
[3]Colorado State University, Fort Collins, Fort Collins, CO 80523, USA

Corresponding author: Zhongyuan Ruan (zyruan@zjut.edu.cn)

ABSTRACT Intelligence transportation system (ITS) and vehicular networks have attracted the research community in the recent years which generate the ''big data'' in traffic. However, the collection and application of the big traffic data is limited by the privacy of people who generate data. Besides, data-driven-based ITS only needs information that could reflect one or more types of vehicles at specific intersections, sections, and road networks, rather than that of each individual vehicle. Overall, intelligent analysis and data fusion of multi-source traffic data play an important role to reduce the phenomenon of privacy disclosure and ensure the quality of data. As a result, a complete method of multi-source traffic data analyzing and processing is proposed in this paper, including the data analysis method based on the spatio-temporal regression model and the data fusion method using evidence theory based on the confidence tensor. Finally, the practical data is used to conform the ways proposed before. And not only do the results show that the implicit privacy information has been removed but also present a higher accuracy of the proceed data.

INDEX TERMS Intelligent transportation system, multi-source traffic data, data analysis, data fusion, data privacy.

I. INTRODUCTION

In general, traffic data can be divided into three parts according to different sources: direct detection data, transport industry data and other related data. The first type data are generated by video checkpoint, microwave detection, coil detection, semaphore and so on; the second type data are from various transport industries, such as GPS data of vehicles, etc.; other related data contain operator data, weather data, public sentiment data and so on [1]. And each of different types of data may exist people's privacy which could be exposed to the public. In-depth analysis and multi-source fusion of various traffic data is the basis of traffic planning and control. And it is also one of the most effective ways to reduce the information of privacy. Various types of data models and comprehensive analysis methods nowadays emerge in an endless stream and show great power and potentiality with the development of information technologies such as artificial intelligence, big data and cloud computing.

Affected by the reliability of various types of detection equipment, the uniqueness of the type of monitoring data and the concerns of privacy disclosure by public opinion [2], [3], the original traffic data often fail to fully satisfy the need for application. Most researchers have focused on the calculation and analysis of traffic assessment indicators, or the preliminary processing of the required traffic parameters from traffic data. There are fewer systematic studies on how to produce high-quality traffic flow characteristics parameters under the circumstance that data consumers do their best to protect data producers' privacy in the early period. So an independent traffic data analysis and processing is necessary for traffic index calculation, signal control and other related applications. The following challenges exist:

(1) The non-formatting and poor data quality seriously restrict the application of data-driven traffic products. It is necessary to formulate uniform formats and standards of data.

(2) The quality of original traffic data is affected by the reliability of the transmission line and the change of detection conditions. The data missing, redundancy, delay and anomalies phenomenon occur frequently. Therefore, how to improving the quality of original single-source data deserves further study.

(3) Although there are more and more data sources, the comprehensive utilization of multi-source data is not obvious enough and the privacy-protection methods is imperfectness. Therefore, it is urgent to study the fusion method of multi-source data.

In this paper, we gave a complete analysis and processing method for multi-source traffic data. Specifically, three aspects of the extraction of traffic characteristic parameters, the abnormal diagnosis and repair of single source data, and the fusion of multi-source data are presented. This method effectively compensates for the poor applied effect caused by the abnormality of the original data and completely removes the privacy information. At the same time, compared with more complex data processing algorithms, this method guarantees the real-time performance of large-scale applications. The experimental results show its effectiveness.

The remainder of the paper is organized as follows: Previous work and researches is presented in Section 2. Section 3 proposes a perfect analysis and processing method based on multi-source spatio-temporal data. Section 4 describes the details of the experiment and shows our results. Finally, the conclusion and farther work are discussed in Section 5.

II. RELATED WORK

The effective analysis and utilization of traffic big data is the key to ITS [4]. However, there are many problems in traffic data, such as diverse types, uneven quality and irregular format, which result in inefficient use of data. Hou *et al.* proposed a newly safety big data computing framework to improve computational efficiency [5]. Kong *et al.* proposed a vehicle movement dataset generation method based on floating car data and functional communities [6]. Min and Wynter presented an algorithm for repairing data, which repaired the missing data by the continuous spatio-temporal data [7]. These works provide an effective method for the processing of single data sources. In practical applications, there are still some limitations in dealing with complex multi-source traffic data. Smith and Demetsky proposed a non-parametric regression prediction method for nonlinear systems in view of the inadequacy of the early prediction model [8]. Hussein studied linear prediction model for freeways [9]. Ouyang *et al.* proposed a combine method based on the characteristics of wavelet transform, particle swarm optimization algorithm and BP neural network for short-term traffic flow forecasting [10]. Wang *et al.* using fog computing to solving offloading problem in IoV systems [11]. Molzahn*et al.* studied the change in traffic flow before traffic breakdown and used traffic data to confirm the microscopic stochastic theory of traffic breakdown developed by Kerner [12]. Xia *et al.* used large-scale data set evaluation to analyze the mobility characteristics in the vehicle's social network [13]. Ning *et al.* established a social-aware group formation framework for information diffusion [14]. Based on the analysis of existing researches and the problems existing traffic big data, a data analysis and processing method including cellular automata data reconstruction, spatio-temporal data diagnosis and repair, and confidence tensor data fusion has been proposed in this paper. Figure 1 shows the overall logic structure of the proposed algorithm. Firstly, according to the relevant knowledge of traffic flow theory, the data is reconstructed through standardization of traffic variables and unification of sampling frequency. Then the spatio-temporal regression model is established and is applied to quality diagnosis, spatio-temporal repair and confidence analysis. Finally, based on the evidence theory, the traffic characteristic parameters with the same physical meaning are fused by multi-source data.

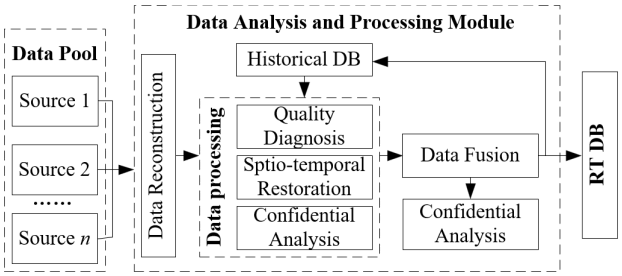


FIGURE 1. Algorithm logic structure diagram.

III. DATA ANALYSIS AND PROCESSING

Once each type of data is uploaded from its source, it enters the data analysis and processing module which includes data reconstruction, multi-dimensional data quality diagnosis, space-time repair and data fusion. Figure 2 shows this process. The processed data is then stored in the database for further mining applications, such as traffic forecasting. This section describes key approaches to data analysis and processing.

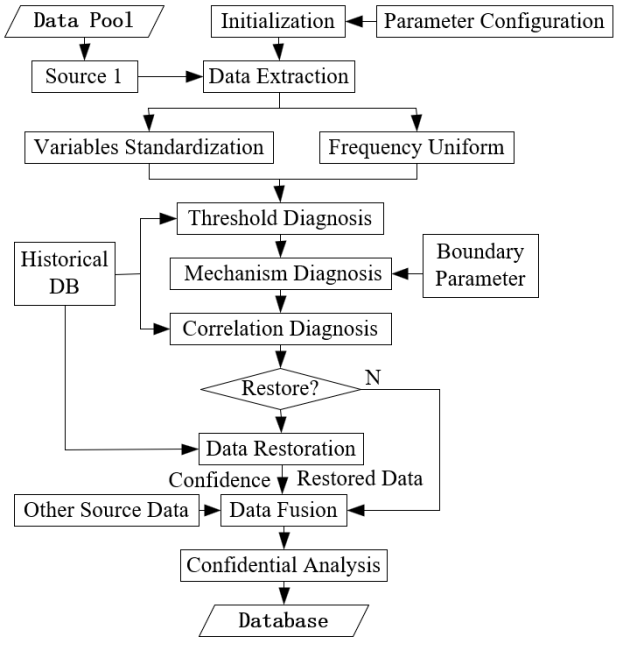


FIGURE 2. Flow chart of traffic data analysis and processing.

A. DATA RECONSTRUCTION

Data reconstruction is the data conversion from one form to another. For multi-source traffic data, it includes the standardization of data format and the unification of sampling frequency.

1) DATA FORMAT STANDARDIZATION

The standardization of data formats refers to a consistent format conversion of data with the same physical meaning from the same or different sources, including dimension unification, value unification, and dimension unification.

Unit unification refers to the process of unifying measurement unit of the same physical meaning data. For example, the unit of speed from video detectors and GPS need convert to m/s or km/h. Value unification is to make the physical value of comparable numerical size for the same data from different sources. For example, the vehicle occupancy time detected by different sizes of detection coils (including video virtual coils) needs to be standardized as lane occupancy rate during green light. The conversion formula is as follows:

$$o_{g,j} = \frac{\sum_{i=1}^{n} t_{j,i}}{t_{g,j}}, \tag{1}$$

where $o_{g,j}$ is the occupancy rate for lane j during the green light in a signal cycle; $t_{j,i}$ is the time in seconds of a detection coil occupied by a vehicle on lane j in phase i being the green signal; n is the number of phase in which vehicles on lane j have the right of passage; $t_{g,j}$ is the green signal duration time for lane j.

The dimension unification refers to the process of converting the detection objects in some data to standard reference. For example, buses, trucks, cars and other vehicles are different in terms of vehicle size, floor area and vehicle speed, and must be converted into a unified standard vehicle. In this paper, the cellular automata model is used to analyze the traffic characteristics of different types of vehicles.

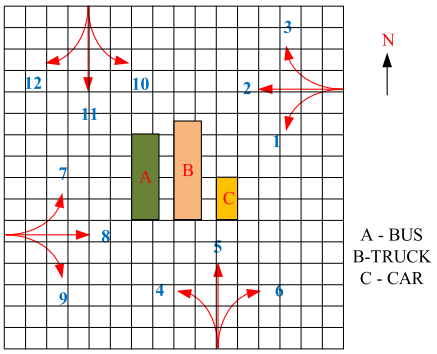


FIGURE 3. Traffic characteristics of different types of vehicle.

Figure 3 shows bus, truck, and car are converted 4:4.5:2 in the spatial dimension. Moreover, in the time dimension, according to the speed of different types of vehicles, the equivalent conversion model can be established. Taking the truck as an example, the conversion model is:

$$q'_{tru} = k_{tru} \cdot q_{tru}, \tag{2}$$

$$k_{tru} = \frac{\gamma \cdot (\frac{l}{v})}{h_0}, \tag{3}$$

where q'_{tru} is the standard equivalent flow of truck in an hour (pcu/h); q_{tru} is the actual flow of track (veh/h); l is the distance between two adjacent trucks in saturated condition (m/veh); v is the average speed of truck through the intersection in saturated condition(m/s); h_0 is the headway of standard vehicle in saturated condition (s/pcu); γ is a correction factor to correct different traffic flow direction or special behavior; k_{tru} is the standard vehicle equivalent conversion factor of trunk. Due to the randomness of traffic operation, a large number of data surveys are usually required to calibrate a reasonable k_{tru}.

2) SAMPLING FREQUENCY UNIFICATION

Traffic data from different sources usually have different output frequencies. For example, the video detection data is uploaded in real time every second; the inductive coil data is output in every phase or cycle; the GPS data of taxi is updated every 15-30 seconds; and the map service provider data is refreshed every 2 minutes or so. Therefore, the data sampling frequency should be unified for the subsequent data processing. Taking the lane flow as an example, the induction coil connected to the signal controller provides the number of vehicle passed q_1 (veh) in the variable signal period T_1 (s) while the video detection device obtains the number of vehicle passed q_2 (veh) in the fixed sampling period T_2 (s). Traffic flow over a unified sampling period ΔT (s) being the minimum common multiple of T_1 and T_2 can be calculated as follows:

$$q'_1 = \frac{\sum_{j}^{n} q_{1,j}}{\Delta T} \times 3600, \quad n = \frac{\Delta T}{T_1}, \tag{4}$$

$$q'_2 = \frac{\sum_{i}^{m} q_{2,i}}{\Delta T} \times 3600, \quad m = \frac{\Delta T}{T_2}, \tag{5}$$

where q'_1 and q'_2 are converted traffic flow at uniform sampling frequency(veh/h).

B. QUALITY DIAGNOSIS

There may be data anomalies in the detection, transmission, and storage, which can be represented by missing data, exceeding the threshold value or deviating from the rationality of the data. Some anomalies are difficult to find though simple filtering. Data anomaly diagnosis has been widely studied all along as the key to the effective implementation of data-driven applications. Kong *et al.* used crowdsourcing technology to analyze long-term traffic anomalies [15]. Ning *et al.* designed a service access system in SIoVs to make a reliability assurance strategy and quality optimization [16]. In this section, taking basic traffic parameters, speed and occupancy as examples, a comprehensive data diagnosis is proposed by threshold method [17], traffic mechanism judgment method [18], and spatial-temporal correlation analysis method.

1) THRESHOLD DIAGNOSIS

a: FLOW THRESHOLD DIAGNOSIS

Each road has a corresponding maximum capacity as a default threshold reflecting road characteristics. In practical application, the threshold is also affected by actual road conditions

and driving habits, which need to be corrected by actual data. Thus, real road capacity C can be expressed as:

$$C = f_c \cdot C_0, \tag{6}$$

where C_0 is the maximum road capacity determined by the road design standard (pcu/h); f_c is the correction factor determined by actual data calibration.

Thus, the reasonable range of traffic flow q is:

$$0 \le q \le C, \tag{7}$$

If q is out of this range, the data will be diagnosed as abnormal which should be repaired.

b: SPEED THRESHOLD DIAGNOSIS

Each road in the city has a speed limit. Most vehicles are driving below the speed limit, but the instantaneous speed of a small number of vehicles will be slightly greater than the speed limit. Thus, the range of the speed threshold is:

$$0 \le v \le f_v \cdot v_{max}, \tag{8}$$

where v is the real speed; v_{max} is the upper-speed limit; f_v is a factor to correct the threshold according data statistics. Similarly, if v is out of this range, the data are considered as abnormal data, which should be repaired.

c: OCCUPATION THRESHOLD DIAGNOSIS

Occupation rate o reflects the density of vehicle in time dimension, which is the ratio of the accumulated time of vehicle occupied detector to a certain observation time T. Its reasonable range is:

$$0 \le o \le 100\%, \tag{9}$$

If o is greater than 100%, the data is considered as abnormal data which needs to be restored.

2) TRAFFIC FLOW MECHANISM DIAGNOSIS

When a data is 0, it is difficult to judge its correctness directly. If other relevant traffic data can be obtained, the data quality can be diagnosed by using its inherent logic mechanism. The mechanism among the three basic parameters of traffic flow, speed and occupancy has been analyzed in Table 1 [19]:

3) SPATIO-TEMPORAL CORRELATION DIAGNOSIS

The threshold and traffic mechanism can be used to diagnose the zero-value anomaly or out-of-bounds anomaly. However, when the data is within the threshold range, it is difficult to determine whether the data is abnormal or not. The data generated by traffic operation is closely related to the traffic supply and traffic demand. The supply of urban roads tends to stabilize in a certain period, and the traffic demand shows periodicity and regularity in the time. Therefore, traffic data is continuous and regular in the time dimension. In the spatial dimension, traffic data of the same section of different lanes or different sections between upstream and downstream of the same road are certainly related to each other.

TABLE 1. Judgment rules on traffic flow mechanism.

No	Condition	Conclusion	Action
1	Being Negative	Error	Delete
2	$q=0, v=0, o=0$	Missing data	Subsequent diagnosis
3	$q\neq 0, v=0, o=0$	Error	Delete
4	$q=0, v\neq 0, o=0$	Error	Delete
5	$q=0, v=0, o\neq 0$	Error	Delete
6	$q=0, v=0, o=1$	Full stop	Subsequent diagnosis
7	$q=0, v\neq 0, o\neq 0$	Error	Delete
8	$q\neq 0, v=0, o\neq 0$	Error	Delete
9	$q\neq 0, v\neq 0, o=0$	Unknown	Subsequent diagnosis
10	$q\neq 0, v\neq 0, o\neq 0$	Unknown	Subsequent diagnosis

Therefore, a multi-dimensional spatio-temporal regression model can be established to further diagnosis the anomalies of the traffic data. The data regression formula is defined as follows:

$$\begin{aligned} D_{reg} = {} & \varphi_1 \cdot [\alpha \cdot D_{his} + (1-\alpha) \cdot D_{pre}] \\ & + \varphi_2 \cdot D_{adj} + \varphi_3 \cdot D_{up} + \varphi_4 \cdot D_{down}, \end{aligned} \tag{10}$$

where D_{reg} is regression mean of data in the section; D_{his} is historical mean value of data in that same period; D_{pre} is historical mean value of data in that forward period; D_{adj} is the data in the same section of adjacent lane in the current period; D_{up} and D_{down} are the data of the upstream and downstream in the current period respectively; $\varphi_1, \varphi_2, \varphi_3, \varphi_4$ are the corrected factors and satisfy $\varphi_1 + \varphi_2 + \varphi_3 + \varphi_4 = 1$.

η is defined as the range of data fluctuation:

$$\eta = \max\left(\frac{D_{max} - D_{med}}{D_{med}}, \frac{D_{med} - D_{min}}{D_{med}}\right), \tag{11}$$

where D_{max}, D_{min} and D_{min} are historical maximum, minimum and median value of data in the section in that same period.

According to formula (10) and (11), if the data satisfies the following conditions, it means the data is normal and can be marked the initial confidence value of 100%. Otherwise, it is marked as abnormal data and need to enter the repair process.

$$D \in [D_{reg} \cdot (1-\eta), D_{reg} \cdot (1+\eta)]. \tag{12}$$

C. SPATIO-TEMPORAL CORRELATION RESTORATION

In view of the correlation of traffic data in two dimensions of time and space, the abnormal data can be repaired by the spatio-temporal correlation restoration method. According to the difference of data quality, the confidence tensor of each type data is defined to evaluate the confidence level of the repaired data. The specific spatio-temporal correlation data repair model is as follows:

$$\begin{aligned} D(k) = {} & \alpha \cdot \frac{1}{m_1} \cdot \sum_{i=1}^{m_1} D_1(k-i) + \beta \cdot \frac{1}{m_2} \cdot \sum_{j=1}^{m_2} D_2(k-j) \\ & + \gamma \cdot \frac{1}{m_3} \cdot \sum_{l=1}^{m_3} D_{3,l}(k), \end{aligned} \tag{13}$$

where $D_1(k-i)$ is the non-trusted real-time data whose initial confidence value is not equal to 100%); m_1 is the number of valid non-trusted real-time data; $D_2(k-j)$ stands for the trusted historical data for the same period; m_2 is the number of valid trusted historical data; $D_{3,l}(k)$ is the trusted data of the adjacent lane of the same type; m_3 is the number of lane providing valid trusted data; α, β and γ the weight factors and satisfy $\alpha+\beta+\gamma=1$.

Due to the differences in the number of trust data in the spatio-temporal dimension, the data quality obtained by the restoration process is not the same. Thus, it is necessary to make a confidence evaluation of the repaired data, which can provide a reference for later application. The confidence analysis of repaired data is divided into two parts: the spatial dimension and the temporal dimension.

The confidence of spatial dimension F_s is defined as follows:

$$F_s = \frac{1-e^{\alpha\cdot m_3}}{1+e^{\alpha\cdot m_3}} \times 100\%, \tag{14}$$

where α is the factor used to correct the consistency between the number of lanes and the confidence level of repaired data.

The confidence of temporal dimension F_t is defined as follows:

$$F_t = [\delta\cdot\frac{1-e^{\beta\cdot m_1}}{1+e^{\beta\cdot m_1}} + (1-\delta)\cdot\frac{1-e^{\gamma\cdot m_2}}{1+e^{\gamma\cdot m_2}}]\times 100\%, \tag{15}$$

where δ is weight coefficient; β and γ are factors used to correct the consistency between the number of the trusted historical data and the confidence level of repaired data.

The maximum value in F_s and F_t is taken as the initial confidence value F of the repaired data:

$$F = \max\{F_s, F_t\}. \tag{16}$$

The confidence tensor of each type data includes three categories: accuracy rate P_a, an error rate P_e, and an uncertainty level P_u, defined as follows:

$$\begin{cases} P_a = F - E_d\cdot\gamma \\ P_e = E_d\cdot\gamma \\ P_u = 1-P_a-P_e, \end{cases} \tag{17}$$

where E_d is the measurement error of each data source devices and is a fixed value; γ is a correction factor related with measurement environment. The final output consists of two parts: one is the repaired data or the original data without repair, and the other is the confidence tensor of the data.

D. MULTI-SOURCE DATA FUSION

Data fusion [20] is an information processing of analyzing and synthesizing data from different sources under certain criteria. Its purpose is to remove data redundancy, and to use multiple data sources to further correct data simultaneously and filter the individual information completely from each single data source. Data fusion can improve the integrity, reliability and accuracy of the data obviously. The method of evidence theory [21], [22] are chosen to select the repaired data from different sources and to fuse the data according to the confidence rules.

Some of the main variables are defined as follows: Set $\mathbf{S}$ represents N data sources of the same class of data having the same recognition framework $\mathbf{\Theta}$ ($\mathbf{\Theta}=\{P_a, P_e, P_u\}$); $n_{\mathbf{\Theta}}$ is the number of target patterns identified within the framework where $n_{\mathbf{\Theta}}=3$; $m_{\mathbf{X}_i}(\mathbf{p}_j)$ is the basic probability assignment of data source combination $\mathbf{X}_i$ to the target pattern $\mathbf{p}_j$ where $\mathbf{X}_i$ and $\mathbf{p}_j$ is the non-empty subset of $\mathbf{S}$ and $\mathbf{\Theta}$. The steps for fusion are as follows:

(1) Calculate all nonempty subsets $\mathbf{X}_i$ in the set $\mathbf{S}$ where $i=1,2,\cdots,2^N-1$ and all $\mathbf{X}_i$ form a set $\mathbf{X}$ representing the possible composition of N data sources. Meanwhile, calculate all nonempty subset $\mathbf{p}_j$ in the set $\mathbf{\Theta}$ where $j=1,2,\cdots,2^{n_{\mathbf{\Theta}}}-1$ and merge $\mathbf{p}_j$ into a set $\mathbf{p}$.

(2) For each data source combination $\mathbf{X}_i$ with unknown confidence, the fusion result of confidence is obtained by formula (18) according to the confidence of each data source:

$$\begin{aligned} m_{\mathbf{x}_i}(\mathbf{p}_j) &= \frac{\sum\limits_{\mathbf{p}_j\cap\mathbf{p}_k=\mathbf{p}_k}\prod\limits_{s\in\mathbf{x}_i} m_s(\mathbf{p}_k)}{\sum\limits_{\mathbf{p}_j\cap\mathbf{p}_k\neq\emptyset}\prod\limits_{s\in\mathbf{x}_i} m_s(\mathbf{p}_k)} \\ &= \frac{\sum\limits_{\mathbf{p}_j\cap\mathbf{p}_k=\mathbf{p}_k}\prod\limits_{s\in\mathbf{x}_i} m_s(\mathbf{p}_k)}{1-\sum\limits_{\mathbf{p}_j\cap\mathbf{p}_k=\emptyset}\prod\limits_{s\in\mathbf{x}_i} m_s(\mathbf{p}_k)}, \end{aligned} \tag{18}$$

where $m_s(\mathbf{p}_k)$ is the known basic probability assignment of one data source $s(s\in\mathbf{X}_i)$ to the target pattern $\mathbf{p}_k$.

(3) Let $\exists\mathbf{X}_1,\mathbf{X}_2\subseteq\mathbf{X}$ and satisfy $m_{\mathbf{X}_1}(\{P_a\}) = \max\{m_{\mathbf{X}_i}(\{P_a\})\}$, $m_{\mathbf{X}_2}(\{P_e\}) = \max\{m_{\mathbf{X}_i}(\{P_e\})\}$. if there is

$$\begin{cases} m_{\mathbf{X}_1}(\{P_a\}) - m_{\mathbf{X}_2}(\{P_e\}) > \varepsilon_1 \\ m_{\mathbf{X}_1}(\{p_u\}) < \varepsilon_2 \\ m_{\mathbf{X}_1}(\{P_a\}) > m_{\mathbf{X}_1}(\{p_u\}), \end{cases} \tag{19}$$

Then $\mathbf{X}_1$ is the final decision combination, where ε_1 and ε_2 are preset thresholds. Otherwise, data fusion cannot be performed through this combination.

(4) Fuse data from the data source in $\mathbf{X}_1$ as follows:

$$D = \sum_{1\le i\le n}\frac{p_{a,i}/(p_{e,i}*p_{u,i})}{\sum\limits_{1\le j\le n} p_{a,j}/(p_{e,j}*p_{u,j})}\dot{D}_i, \tag{20}$$

Where n is the number of data sources of $\mathbf{X}_1$.

IV. EXPERIMENT

The experiment selects video data from a main road and GPS data from AMAP, a map service provider, in Xiaoshan, China, on 30 September 2017 6: 00-21: 00. The observation time is divided into 180 small segments with an interval of 5 minutes and the average speed of each segment from two data sources is obtained sequentially. Taking the AMAP speed as an example, the data spatio-temporal anomaly identification method is used to identify the abnormal speed. At the same time, a number of speeds are randomly removed and marked as data loss. Then the abnormal data and the missing data are repaired. Finally, the repaired AMAP speed is fused with the

video speed processed in the same way, and the effectiveness of the proposed method is verified by comparing the confidence tensor before and after fusion.

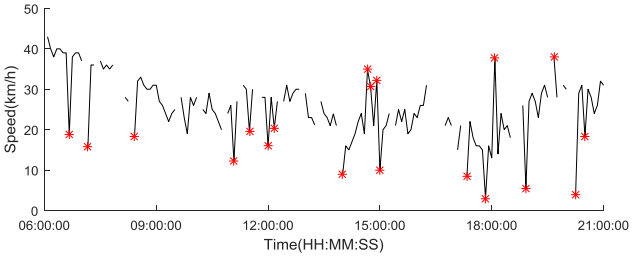


FIGURE 4. Spatiotemporal correlation diagnosis results of AMAP speed.

The results of anomaly diagnosis of AMAP speed are shown in Figure 4. In the figure, there are unreasonable tempo mutations in some segments marked with '$*$' in red, and these speeds are diagnosed as abnormal data. The reason for the anomaly may be that the sample number of vehicles passing through the road section during this time interval is less or abnormal parking. In addition, the discontinuity of the curve indicates the existence of data loss.

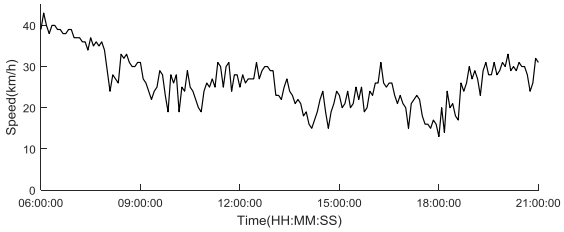


FIGURE 5. Repair results of AMAP speed.

The repair results of the abnormal and missing data from AMAP speed are shown in Figure 5. From the comparison of the repaired curve in Figure 5 and the original curve in Figure 4, it shows that the repaired data tend to be smoother than the abnormal data, and the repair results for randomly deleted data are basically consistent with adjacent data on the order of magnitude and the trend of variation. From the point of view of error quantification, the mean error of repaired data is 3.54km/h, and the mean error rate is less than 8.73%. It can be seen that the spatial-temporal correlation diagnosis and repair method can effectively diagnose and repair anomaly data, and the repaired data can meet the needs of subsequent related applications.

During the fusion progress, we filter the privacy data from each data source and picked up the common ones. The fusion result of the repaired AMAP speed and video speed is shown in Figure 6. Meanwhile, the accuracy rate, error rate and uncertainty of the data confidence tensor before and after each fusion are selected as the index, and the confidence comparison result in multiple dimensions are shown in Figures 7–9.

From figure 6, it shows that the overall speed curve has a tendency of being high only in the morning and at night, but lower in the daytime, which meets the traffic characteristics

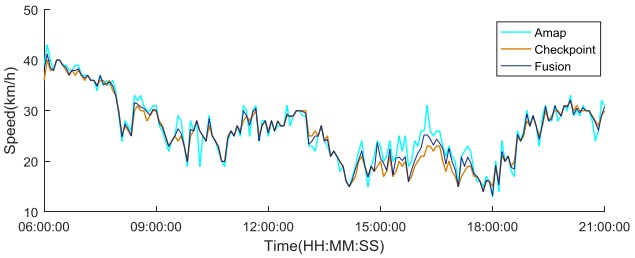


FIGURE 6. The speed fusion result of AMAP and video.

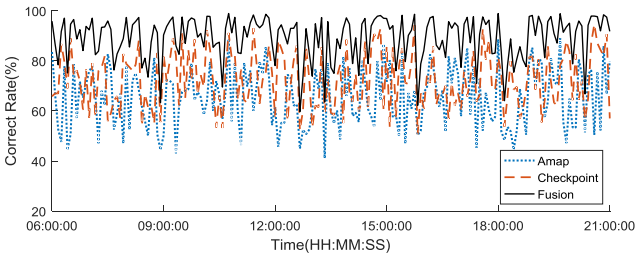


FIGURE 7. comparison Accuracy rate before and after fusion.

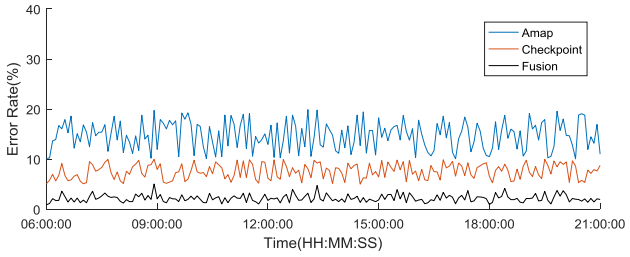


FIGURE 8. Comparison of error rates before and after fusion.

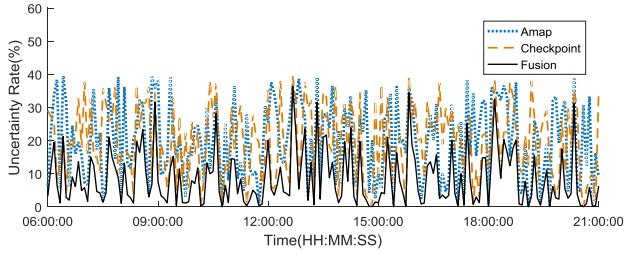


FIGURE 9. Comparison of uncertainty before and after fusion.

of people returning home on holiday before National Day. From Figure 7-9, it can be seen that the accuracy rate of data before fusion is generally low, while the error rate and uncertainty are relatively high, and the accuracy rate of data after fusion is obviously improved, and the error rate and data uncertainty are also obviously reduced. Therefore, the fusion of multi-source data can greatly improve the accuracy and availability of data.

V. CONCLUSION

Depth analysis and multi fusion of traffic big data has become a hot and difficult point in data-driven applications. This paper presents an analysis and processing method of data reconstruction, spatio-temporal diagnosis and correlation restoration and data fusion based on multi-source traffic data. Taking the speed as an example, the above method has

been verified. The experiment results show that this method can effectively improve integrity and quality of the data and has an important value for the comprehensive application of multi-source traffic data.

This paper adopts the traditional mathematical method to analyze and deal with traffic big data. With the development of artificial intelligence, new data processing models and methods are emerging, and their applications in engineering will be shown its tremendous powers and huge potentials.

REFERENCES

[1] H.-P. Lu, Z.-Y. Sun, and W.-C. Qu, "Big data and its applications in urban intelligent transportation system," *Transp. Syst. Eng. Inf.*, vol. 15, no. 5, pp. 45–52, May 2015, doi: 10.16097/j.cnki.1009-6744.2015.05.007.

[2] D. Y. Bao and J. Wang, "Research on cleaning and repairing methods of vehicle detector abnormal data," *Internet Things Technol.*, vol. 5, no. 10, pp. 82–83 and 86, Oct. 2015, doi: 10.16667/j.issn.2095-1302.2015.10.005.

[3] H. Zhang, S. Liu, W. Jiao, D. Li, and Y. Liu, "Device abnormality detection method based on variable threshold information detector," *J. Mech. Eng.*, vol. 49, no. 8, pp. 25–31, Mar. 2013, doi: 10.3901/JME.2013.08.025.

[4] Z. Xiao, X. Fu, L. Zhang, L. Ponnambalam, and R. S. M. Goh, "Data-driven multi-agent system for maritime traffic safety management," in *Proc. IEEE 20th Int. Conf. Intell. Transp. Syst. (ITSC)*, Yokohama, Japan, Oct. 2017, pp. 1–6.

[5] W. G. Hou, Z. Ning, L. Guo, and X. Zhang, "Temporal, functional and spatial big data computing framework for large-scale smart grid," *IEEE Trans. Emerg. Topics Comput.*, to be published, doi: 10.1109/TETC.2017.2681113.

[6] X. J. Kong *et al.*, "Mobility dataset generation for vehicular social networks based on floating car data," *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 3874–3886, May 2018, doi: 10.1109/TVT.2017.2788441.

[7] W. Min and L. Wynter, "Real-time road traffic prediction with spatio-temporal correlations," *Transp. Res. C Emerg. Technol.*, vol. 19, no. 4, pp. 606–616, 2011, doi: 10.1016/j.trc.2010.10.002.

[8] B. L. Smith and M. J. Demetsky, "Traffic flow forecasting: Comparison of modeling approaches," *J. Transp. Eng.*, vol. 123, no. 4, pp. 261–266, Jul. 1997, doi: 10.1061/(ASCE)0733-947X(1997)123:4(261).

[9] H. Dia, "An object-oriented neural network approach to short-term traffic forecasting," *Eur. J. Oper. Res.*, vol. 131, no. 2, pp. 253–261, Jun. 2001, doi: 10.1016/S0377-2217(00)00125-9.

[10] L. Ouyang, F. Zhu, G. Xiong, H. Zhao, F. Wang, and T. Liu, "Short-term traffic flow forecasting based on wavelet transform and neural network," in *Proc. IEEE 20th Int. Conf. Intell. Transp. Syst. (ITSC)*, Yokohama, Japan, Oct. 2017, pp. 1–6.

[11] X. Wang, Z. Ning, and L. Wang, "Offloading in Internet of vehicles: A fog-enabled real-time traffic management system," *IEEE Trans. Ind. Informat.*, to be published, doi: 10.1109/TII.2018.2816590.

[12] S.-E. Molzahn, B. S. Kerner, H. Rehborn, S. L. Klenov, and M. Koller, "Analysis of speed disturbances in empirical single vehicle probe data before traffic breakdown," *IET Intell. Transp. Syst.*, vol. 11, no. 9, pp. 604–612, Nov. 2017, doi: 10.1049/iet-its.2016.0315.

[13] F. Xia, A. Rahim, X. Kong, M. Wang, Y. Cai, and J. Wang, "Modeling and analysis of large-scale urban mobility for green transportation," *IEEE Trans. Ind. Informat.*, vol. 14, no. 4, pp. 1469–1481, Apr. 2018, doi: 10.1109/TII.2017.2785383.

[14] Z. Ning, X. Wang, X. Kong, and W. Hou, "A Social-aware group formation framework for information diffusion in narrowband Internet of Things," *IEEE Internet Things J.*, vol. 5, no. 3, pp. 1527–1538, Jun. 2018, doi: 10.1109/JIOT.2017.2777480.

[15] X. Kong, X. Song, F. Xia, H. Guo, J. Wang, and A. Tolba, "LoTAD: long-term traffic anomaly detection based on crowdsourced bus trajectory data," *World Wide Web*, vol. 21, no. 3, pp. 825–847, Aug. 2017, doi: 10.1007/s11280-017-0487-4.

[16] Z. Ning, X. Hu, Z. Chen, M. Zhou, B. Hu, and J. Chen, "A cooperative quality-aware service access system for social Internet of vehicles," *IEEE Internet Things J.*, vol. 5, no. 4, pp. 2506–2517, Aug. 2018, doi: 10.1109/JIOT.2017.2764259.

[17] H. P. Lu, Z. Y. Shun, and W. C. Qu, "Dynamic threshold based real-time traffic flow fault data identification method," *Civil Eng. J.*, vol. 48, no. 11, pp. 126–132, Nov. 2015, doi: 10.15951/j.tmgcxb.2015.11.017.

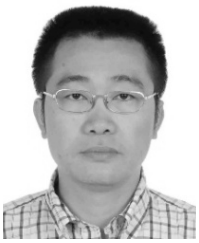
[18] H. Han, "Analysis of traffic flow mechanism based on data mining," M.S. thesis, Dept. Trans. Eng., South China Univ. Technol., Guangzhou, China, 2014.

[19] C. Xu, Z. Qu, P. Tao, and S. Jin, "Real-time screening and recovery of abnormal values of dynamic traffic data," *J. Harbin Eng. Univ.*, vol. 37, no. 2, pp. 211–217, Apr. 2016, doi: 10.11990/jheu.201503045.

[20] N.-E. El Faouzi and L. A. Klein, "Data fusion for ITS: Techniques and research needs," in *Proc. Int. Symp. Enhancing Highway Perform. (ISEHP)*, Berlin, Germany, 2016, pp. 495–512.

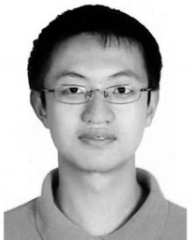
[21] W. Li, "Research and application of multi-sensor fusion algorithm based on DS evidence theory," M.S. thesis, Dept. Electron. Eng., Taiyuan Univ. Technol., Taiyuan, China, 2015.

[22] Y. Lemeret and E. Lefevre, "Evidence theory for data fusion in transportation systems," *IFAC Proc. Volumes*, vol. 37, no. 19, pp. 81–86, Oct. 2004, doi: 10.1016/S1474-6670(17)30663-8.



GUOJIANG SHEN received the B.Sc. degree in control theory and control engineering and the Ph.D. degree in control science and engineering from Zhejiang University, Hangzhou, China, in 1999 and 2004, respectively.

He is currently a Professor with the College of Computer Science and Technology, Zhejiang University of Technology. His current research interests include artificial intelligence, theory, big data analytics, and intelligent transportation system.



XIAO HAN was born in Hangzhou, Zhejiang, China, in 1996. He received the B.Sc. degree from the Zhejiang University of Technology, China, in 2018, where he is currently pursuing the M.S. degree.

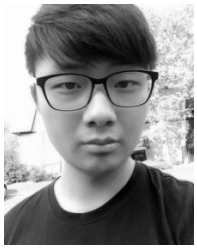
He participated in the City Brain Project in 2017. His current research interests include big data analytics and intelligent transportation system.

JUNJIE ZHOU received the B.Sc. degree in control engineering from Zhejiang University, Hangzhou, China, in 2016.

He is currently an Algorithmic Engineer with Zhejiang Supcon Information Co., Ltd., Hangzhou. His current research interests include traffic signal control algorithms, big data analytics, and intelligent transportation system.

ZHONGYUAN RUAN received the B.Sc. degree in physics from Guizhou University and the Ph.D. degree in physics from East China Normal University, Shanghai, China, in 2008 and 2013, respectively.

He is currently a Lecturer with the College of Computer Science and Technology, Zhejiang University of Technology. His current research interests include complex systems and complex networks.



QIHONG PAN received the bachelor's degree in software engineering from the School of Software Engineering, Beijing Jiaotong University, Beijing, China, in 2015. He is currently pursuing the master's degree in electrical and computer engineering from Colorado State University, Fort Collins, CO, USA. His research interests include big data and machine learning.